Classification and automatic annotation extension of images using Bayesian network

S. Barrat[1], S. Tabbone[1]

LORIA-UMR 7503, University of Nancy 2,
BP 239, 54506 Vandœuvre-lès-Nancy, France
{barrat,tabbone}@loria.fr

Abstract. In many vision problems, instead of having fully annotated training data, it is easier to obtain just a subset of data with annotations, because it is less restrictive for the user. For this reason, in this paper, we consider especially the problem of classifying weakly-annotated images, where just a small subset of the database is annotated with keywords. In this paper we present and evaluate a new method which improves the effectiveness of content-based image classification, by integrating semantic concepts extracted from text, and by automatically extending annotations to the images with missing keywords. Our model is inspired from the probabilistic graphical model theory: we propose a hierarchical mixture model which enables to handle missing values. Results of visual-textual classification, reported on a database of images collected from the Web, partially and manually annotated, show an improvement by 32.3% in terms of recognition rate against only visual information classification. Besides the automatic annotation extension with our model for images with missing keywords outperforms the visual-textual classification by 6.8%. Finally the proposed method is experimentally competitive with the state-of-art classifiers.

Keywords: probabilistic graphical models, Bayesian networks, image classification, image annotation

1 Introduction

The rapid growth of internet and multimedia information has shown a need in the development of multimedia information retrieval techniques, especially the image retrieval. We can distinguish two main trends.
The first one, called "text-based image retrieval", consists in applying text-retrieval techniques from fully annotated images. The text describes high-level concepts but this technique requires a tedious work of annotation. The second approach, called "content-based image retrieval" is a more young field. These methods rely on visual features (color, texture or shape) computed automatically, and retrieve images using a similarity measure. However the obtained performances are not really acceptable, except in the case of well-focused corpus.

In order to improve the recognition, a solution consists in combining visual and semantic informations. Some researchers have already explored this possibility [1–4].
Finally, automatic image annotation can be used in image retrieval systems to organize and locate images of interest from a database, or to perform visual-textual classification. This method can be regarded as a type of multi-class image classification with a very large number of classes, as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are used by machine learning techniques to attempt to automatically apply annotations to new images. Many works have been proposed in this sense [5–10].
In this perspective, the contribution of this paper is to propose a scheme for image classification optimization, by using a joint visual-text clustering approach and automatically extending image annotations.

2 Motivations

The motivation of the model presented here is to use it for the both tasks of weakly-annotated image classification and annotation. In fact the classification methods before mentionned are efficient but they requires that all image, or image blobs are annotated. Moreover the annotation models are not enable to classify images.
The proposed model does not require that all images be annotated: when an image is weakly annotated, the missing keywords are considered as missing values. Besides it enables to automattically extend existing annotations to weakly-annotated images, without User intervention.
The model [5] is the most related to our approach, because it enables to classify images based on visual and textual features and to automatically annotate new images. However our model is less restrictive for the user. In fact our classifier does not need that all images be annotated, and the existing keywords are associated to the whole images, not to image regions.
The proposed approach is derived from the probabilistic graphical model theory. We introduce a method to deal with missing data in the context of text annotated images as defined in [5, 4]. The uncertainty around the association between a set of keywords and an image is tackled by a joint probability distribution over the dictionary of keywords and the numerical features extracted from our collection of grey-level and color images.
The rest of this paper is organized as follows: Section 3 decribes the probabilistic model of weakly-annotated image representation and how to use it to classify and extend existing annotations to images. Section 4 presents the experimental results. Finally conclusions and future works are given in Section 5.

3 Representation and classification of weakly-annotated images

Our work is focused on weakly-annotated image modelling and classification. Now visual descriptors often provide vectors of continuous values, and the associated keywords often correspond to discrete variables. So we have chosen to construct a Bayesian classifier which allows for the discrete and continuous variable combination and the problem of missing values handling.

Let f_j be a query image characterized by a set of features F. F is composed of:

- m visual features, denoted $v_1, ..., v_m$,
- n possible keywords, denoted $KW\ 1, ..., KW\ n$.

The chosen visual features are issued from one color descriptor, a color histogram [11], and one shape descriptor based on the Fourier/Radon transform [12]. We are interested in the probability distributions of these features and their conditional dependence relations. Let us consider the visual features as continuous random variables and theirs associated keywords as discrete variables. This model is too big to be represented as a unique joint probability distribution, therefore it is required to introduce some sparse and structural *a priori* knowledge. The probabilistic graphical models, and especially Bayesian networks, are a good way to solve this kind of problem. In fact within Bayesian networks the joint probability distribution is replaced by a sparse representation only among the variables directly influencing one another. Interactions among indirectly-related variables are then computed by propagating inference through a graph of these direct connections. Consequently the Bayesian networks are a simple way to represent a joint probability distribution over a set of random variables, to visualize the conditional properties and to compute complex operations like probability learning and inference, with graphical manipulations. Then a Bayesian network seems to be appropriate to represent and classify images and associated keywords.

3.1 A Gaussian-Mixtures and Multinomial mixture model

We present a hierarchical probabilistic model of multiple-type data (images and associated keywords) in order to classify large annotated image databases. A Gaussian Mixtures and Multinomial Mixture model is proposed. In fact, the observation of some peaks on the different histograms of the feature variables, has led us to consider that the visual features can be estimated by mixtures of Gaussian densities. The discrete variables corresponding to the possible keywords are assumed to be distributed as a multinomial distribution over the vocabulary of keywords.

Now let F be the training set composed of m instances $f_{1_i}, ..., f_{m_i}, \forall i \in \{1, ..., n\}$, where n is the dimension of the signatures provided by the concatenation of the feature vectors issued from the computation of all the descriptors on each image on the training set. Each instance $f_j, \forall j \in \{1, ..., m\}$ is then

characterized by n continuous variables. A supervised classification is considered then F instances are divided into k classes $c_1, ..., c_k$. Let $G_1, ..., G_g$ be g groups whose each has a Gaussian density with a mean $\mu_l, \forall l \in \{1, ..., g\}$ and a covariance matrix $\sum_l$. Besides, let $\pi_1, ..., \pi_g$ be the proportions of the different groups, $\theta_l = (\mu_l, \sum_l)$ the parameter of each Gaussian and $\Phi = (\pi_1, \pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ the global mixture parameter. Then the probability density of F conditionally to the class $c_i, \forall i \in \{1, ..., k\}$ can be defined by

$$P(f, \Phi) = \sum_{l=1}^{g} \pi_l p(f, \theta_l)$$

where $p(f, \theta_l)$ is the multivariate Gaussian defined by the parameter θ_l.

Then, we have one Gaussian Mixture Model per class. This problem can be represented by the probabilistic graphical model in Figure 1, where:

- The "Class" node is a discrete node, which can take k values corresponding to the pre-defined classes $c_1, ..., c_k$.
- The "Component" node is a discrete node which corresponds to the components (i.e the groups $G_1, ..., G_g$) of the mixtures. This variable can take g values, i.e the number of Gaussians used to compute the mixtures. It's an hidden variable which represents the weight of each group (i.e the $\pi_l, \forall l \in \{1, ..., g\}$).
- The "Gaussian" node is a continuous variable which represents each Gaussian $G_l, \forall \in \{l = 1, ..., g\}$ with its own parameter ($\theta_l = (\mu_l, \sum_l)$). It corresponds to the set of feature vectors in each class.
- Finally the edges represent the effect of the class on each Gaussian parameter and its associated weight. The green circle is just used to show the relation between the proposed probabilistic graphical model and GMMs: we have one GMM (encircled in green), composed of Gaussians and their associated weight, per class.

Now the model can be completed by the discrete variables, denoted $KW\ 1, ..., KW\ n$, corresponding to the possible keywords associated to the images. Dirichlet priors [13], have been used for the probability estimation of the variables $KW\ 1, ..., KW\ n$. That is we introduce additional pseudo counts at every instance in order to ensure that they are all "virtually" represented in the training set. Therefore every instance, even if it is not represented in the training set, will have a not null probability. Like the continuous variables corresponding to the visual features, the discrete variables corresponding to the keywords are included in the graphical model by connecting them to the class variable.

Then our classifier can be depicted by the Figure 2. The hidden variable "α" shows that a Dirichlet prior is used. The box around the variable KW denotes n repetitions of KW, for each keyword. This Bayesian classifier means that each image and its keywords are assumed to have been generated conditional on the same class. Therefore the resulting multinomial and Gaussian mixture parameters should correspond: concretely if an image, represented by visual descriptors, has an high probability under a certain class, then its keywords should have an high probability under the same class.

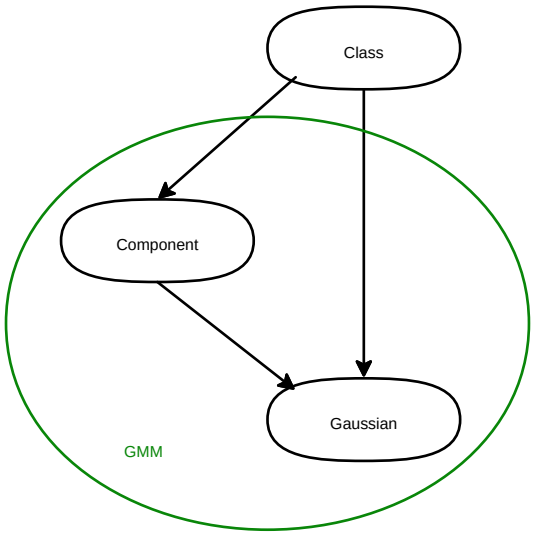

Fig. 1. A Probabilistic graphical model as GMMs

3.2 Parameter learning and inference

The EM algorithm has been used to learn the Gaussian mixture parameters. But our major problem deals with missing values. Indeed only some data are fully observed. It's the case of all visual features for color images or just shape features for grey-level images. On the contrary the color features for grey-level images, and especially some keywords for a large subset of images, are missing. Concerning the visual features, the missing values are clearly homogeneously distributed (because they correspond to grey-level images). But the missing values are randomly distributed for the variables $KW\ i$, $\forall i \in \{1, ..., n\}$. This kind of problem can be tackled with the EM algorithm too. The general purpose of this algorithm, explained in detail in [14] consists in computing, in an iterative way, the likelihood maximum when the instances can be viewed as incomplete data: each iteration consists of an Expectation computation step before a Maximization step. This algorithm has been chosen for its simplicity and generality. Applied to the parameter learning problem, the EM algorithm starts off with the random model initialisation, then, at each iteration, the parameters are updated in order to obtain a local optimal of the maximum-likelihood. An inference algorithm is required to compute the expected sufficient statistics in the expectation step. Then, in the maximization step, the parameters of the model are adjusted to fit the data.

An inference algorithm is also necessary to classify new images. Indeed, the inference process consists in computing posterior probability distributions of one or several other subsets of nodes. In the case of classification, the class node is inferred. According to our Bayesian network topology, the inference process propagates the values from the image feature level represented by the "Gaussian"

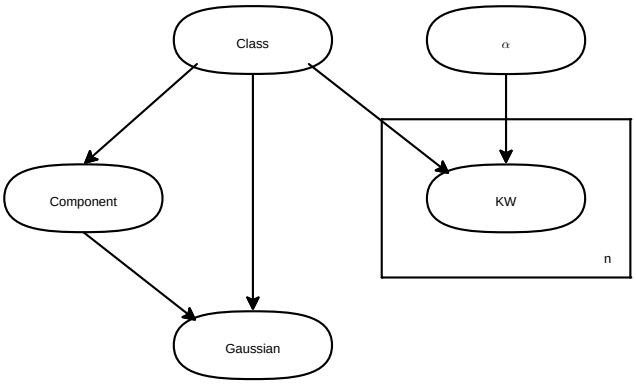

Fig. 2. The Gaussian-Mixtures and Multinomial mixture model

node, trough the "Component" and Keyword nodes, until the "Class" node level. A message passing algorithm [15] is applied to the network. In this technique, each node is associated to a processor, which can send some messages to its neighbors, in an asynchronous way, until it reaches a stability.

Thus a query image f_j, characterized by its visual features $v_{j_1}, ..., v_{j_m}$ and its possible keywords $KW\ 1_j, ..., KW\ n_j$ is considered as an "evidence" represented by:

$$P(f_j) = P(v_{j_1}, ..., v_{j_m}, KW\ 1_j, ..., KW\ n_j) = 1$$

when the network is evaluated. Thanks to the inference algorithm, the probabilities of each node are updated in function of this evidence. After the belief propagation, we know, $\forall i \in \{1, ..., k\}$, the posterior probability $P(c_i|f_j) = P(c_i|v_{j_1}, ..., v_{j_m}, KW\ 1_j, ..., KW\ n_j)$. The query f_j is assigned to the class c_i which maximizes this probability.

3.3 Annotation extension of images

Given an image without keyword, or a weakly annotated image, the proposed Bayesian model before described can be used to compute a distribution over words conditionally to the image and its possible existing keywords. In fact, for a query image f_j annotated by $k, \forall k \in \{0, ..., n\}$ keywords, where n is the maximum keyword number per image, the inference algorithm enables to compute the posterior probability $P(KW_{i_j}|f_j, KW_1, ..., KWk), \forall i \in \{k+1, ..., n\}$. This distribution represents a prediction of the missing keywords for that image.

For example, let us consider Table 1 which presents 3 images with possible keywords and the keywords obtained after automatic annotation extension. The first image, without annotation, has been annotated by 2 suitable keywords. In the same way, the second image annotation, composed at the beginning of 2 keywords, has been extended to 3 keywords. The added keyword, "sunset" is

image	initial possible keywords	keywords after annotation extension
		bridge water
	bridge cloud	bridge cloud sunset
	bridge	bridge cloud sunset

Table 1. Examples of images and possible keywords before and after annotation extension

suitable. On the contrary, the third image, initially annotated by 1 keyword, has obtained 2 other keywords after the automatic extension. But the second new keyword, "sunset" is wrong. This mistake is probably due to the large number of database images annotated by these 3 keywords "bridge", "cloud" and "sunset", and the inference algorithm.

4 Experimental results

In this section, we present an evaluation of our model on more than 3000 free images collected from the Web, and kindly provided by Kherfi et al. [4]. These images are split up into 16 classes. For example, 4 images of the class "horse" are given in Figure 3.

Fig. 3. Examples of horse-class images

65% of the image database have been manually annotated by 1 keyword, 28% by 2 keywords and 6% by 3 keywords, using a vocabulary set of 39 keywords. For example, among the 4 images from the Figure 3, the first image is annotated by the 2 keywords "animal" and "horse". The second is only annotated by one keyword: "animal". The others have not been annotated. We have evaluated our method by performing 5 cross validations whose each proportion of the training set is 25%, 35%, 50%, 65% and 75% of the database, the remaining respectively 75%, 65%, 50%, 35% and 25% are hold for test set. In each case the tests are repeated 10 times in order that each database instance would be used for the

training and the test. For each training set size, the recognition rate is obtained by taking the mean recognition rate of the 10 tests. Since we want to improve the recognition rate by introducing semantics, and automatically extending image annotations, we limit ourselves to the experiments comparing:

- the visual-textual classification to the only visual information classification,
- the visual-textual classification before and after automatic annotation extension of all images,
- the proposed model of visual-textual classification to two state-of-art classifiers

Let us consider Table 2. The notation "C + S" means that the color and shape descriptors ("C" for Color, "S" for Shape) have been combined. The notation "C + S + KW" indicates that both visual descriptors and textual information ("KW" for Keywords) have been combined. The recognition rates confirm that combining visual with semantic features performs always better than any of them alone. In fact we observe that the combination of our visual features and keywords (when they are available) increases the recognition rate by 38.5% compared to the results of color information alone, 58% compared to the shape information classification and 37% compared to the only textual information classification. Besides we can notice that for all experiments, combining the both visual descriptors is better, by 16% on average, to use just one of them. Finally, the visual-textual classification shows an improvement by 32.3% in terms of recognition rate against only visual information classification.

Specifications		Color	Shape	Keywords	C + S	C + S + KW
training part	test part					
25%	75%	35	17.8	36.6	39.4	69.7
35%	65%	36.9	18.1	38.9	42.2	74.4
50%	50%	38.7	18.5	41.1	45	79.1
65%	35%	41.1	20.6	41.5	46.6	81.7
75%	25%	43.5	21.8	45.1	52.9	82.9

Table 2. Recognition rates (in %), obtained by the GM-M model, of only visual classification vs. visual-textual classification

Then, Table 3 shows the effectiveness of our approach (GM-M mixture) compared to the SVM and FKNN classifiers. The results have been obtained by using the visual features and theirs possible associated keywords. It appears that the GM-M mixture results are always better than the ones of SVM and FKNN results.

Finally, annotations have been extended to all images of the database in order that each would be annotated by 3 keywords. Then, to evaluate the quality of this annotation extension, the visual-textual classification has been re-done, with the same specifications as in Table 2. Table 4 shows the efficiency of our automatic annotation extension. In fact, the recognition rates after automatic annotation

Specifications		SVM	FKNN $k=1$	FKNN $k=m$	GM-M mixture
training part	test part				
25%	75%	38.3	59.1	58.5	69.7
35%	65%	41.3	62.3	58.3	74.4
50%	50%	39.9	68.2	58.2	79.1
65%	35%	40.5	72.9	67	81.7
75%	25%	41.9	73.2	69.3	82.9

Table 3. Recognition rates (in %) of the SVM classifier and the FKNN vs. our Gaussian-Mixtures and multinomial mixture model

extension are always better than before. Moreover the automatic annotation extension outperforms the recognition by 6.8% on average.

Specifications		Before annotation extension	After annotation extension
training part	test part		
25%	75%	69.7	77
35%	65%	74.4	79.3
50%	50%	79.1	85.4
65%	35%	81.7	87.6
75%	25%	82.9	92.7

Table 4. Recognition rates (in %), obtained by the GM-M model, of visual-textual classification before and after automatic annotation extension

5 Conclusion and future works

We have proposed an efficient model which enables to combine visual and textual information, handle missing values and extend image annotations to other images. Our experiments have been done on a partially annotated Web image database. The results show that our visual-textual classification method improves the recognition rate compared to the only visual information classification. Moreover our Bayesian network can be used to extend image annotations, what outperforms the recognition rate. Finally the proposed method is competitive with state-of-art classifiers. Further works will be devoted to capture the user's preference by considering a relevance feedback process. More precisely, the user's preference can be represented by the network parameter update (i.e the probabilities of each variable in function of the new classified instance) during the inference process.

References

1. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D.M., Jordan, M.I.: 2003, matching words and pictures. Journal of Machine Learning Research **3**(6)

(2003) 1107–1135
2. Benitez, A., Chang, S.F.: Perceptual knowledge construction from annotated image collections. ICME '02 **1** (2002) 189–192
3. Grosky, W.I., Zhao, R.: Negotiating the semantic gap: From feature maps to semantic landscapes. In: SOFSEM '01. (2001) 33–52
4. Kherfi, M.L., Brahmi, D., Ziou, D.: Combining visual features with semantics for a more effective image retrieval. In: ICPR '04. Volume 2. (2004) 961–964
5. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: SIGIR '03. (2003) 127–134
6. Gao, Y., Fan, J., Xue, X., Jain, R.: Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. In: ACM MULTIMEDIA '06. (2006) 901–910
7. Yang, C., Dong, M., Hua, J.: Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: CVPR '06. (2006) 2057–2063
8. Wang, C., Jing, F., Zhang, L., Zhang, H.J.: Image annotation refinement using random walk with restarts. In: ACM MULTIMEDIA '06. (2006) 647–650
9. Rui, X., Li, M., Li, Z., Ma, W.Y., Yu, N.: Bipartite graph reinforcement model for web image annotation. In: ACM MULTIMEDIA '07. (2007) 585–594
10. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. CVPR '04 **2** (2004) 1002–1009
11. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1) (1991) 11–32
12. Tabbone, S., Wendling, L.: Technical Symbols Recognition Using the Two-dimensional Radon Transform. In: ICPR'02. Volume 2. (2002) 200–203
13. Robert, C.: A decision-Theoretic Motivation. Springer-Verlag (1997)
14. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1) (1977) 1–38
15. Kim, J.H., Pearl, J.: A computational model for combined causal and diagnostic reasoning in inference systems. In: IJCAI-83. (1983) 190–193